June 2026

WORKING PAPER SERIES

2026-EQM-05

Measuring Operational Inefficiency in Urban Road Traffic Systems Using a By-Production Technology: A Machine Learning Approach Based on Convex and Nonconvex Efficiency Analysis Trees

Xiaoqing CHEN
The Research Institute for Risk Governance and Emergency Decision-Making, School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, Jiangsu, China. 003694@nuist.edu.cn
School of Management, Hefei University of Technology. Hefei, Anhui, 230009, China.

Kristiaan KERSTENS
Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille ´Economie Management, F-59000 Lille, France. k.kerstens@ieseg.fr. Tel: +33 320545892 (switchboard), Fax: +33 320574855

Shuqi XU
Corresponding author. College of Management, Hefei University, Hefei, Anhui, 230601, China. xusq@hfuu.edu.cn

Shirong ZHAO
Co-corresponding author. School of Finance, Dongbei University of Finance and Economics, Dalian, 116025, Liaoning, China. shironz@163.com

IÉSEG School of Management Lille Catholic University 3, rue de la Digue F-59000 Lille Tel: 33(0)3 20 54 58 92
www.ieseg.fr

Measuring Operational Inefficiency in Urban Road Traffic Systems Using a By-Production Technology: A Machine Learning Approach Based on Convex and Nonconvex Efficiency Analysis Trees

Xiaoqing Chen [*†] Kristiaan Kerstens[‡] Shuqi Xu[§] Shirong Zhao[¶]

June 18, 2026

Abstract

This paper challenges a central but rarely tested assumption in the by-production framework for modelling technologies that jointly generate desirable and undesirable outputs: the convexity of its constituent sub-technologies. We argue that imposing convexity without empirical validation may lead to misspecified frontiers and biased efficiency assessments. To address this gap, we employ a data-driven methodology that integrates traditional nonparametric frontier estimators with machine-learning based Efficiency Analysis Trees (EAT) and its convexified extension (CEAT). This combination allows us to flexibly approximate the production frontier under alternative geometric assumptions and, for the first time in the by-production literature, to statistically test convexity for each sub-technology using the Kneip-Simar-Wilson (KSW) procedure. Applying the approach to a panel of 119 Chinese cities over the period 2014-2021 and distinguishing between desirable outputs and undesirable outputs, we find a notable asymmetry: convexity is systematically rejected for the desirable-output sub-technology, whereas for the undesirable-output sub-technology the evidence is mixed, with convexity rejected in some years but not in others. These results caution against the routine imposition of convexity in by-production models and provide a statistical foundation for hybrid estimation strategies. Moreover, we show that incorrect convexity assumptions produce substantial divergences in efficiency distributions, underscoring the practical relevance of testing and method choice. The proposed framework offers a robust tool for operational efficiency analysis in urban road traffic systems and other settings where joint production of desirable and undesirable outputs is present.

KEYWORDS: Operational inefficiency; Free disposal hull; Data envelopment analysis; Regression trees; Urban road traffic systems.

[*]The Research Institute for Risk Governance and Emergency Decision-Making, School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, Jiangsu, China. 003694@nuist.edu.cn

[†]School of Management, Hefei University of Technology. Hefei, Anhui, 230009, China.

[‡]Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, F-59000 Lille, France. k.kerstens@ieseg.fr. Tel: +33 320545892 (switchboard), Fax: +33 320574855

[§]Corresponding author. College of Management, Hefei University, Hefei, Anhui, 230601, China. xusq@hfuu.edu.cn

[¶]Co-corresponding author. School of Finance, Dongbei University of Finance and Economics, Dalian, 116025, Liaoning, China. shironz@163.com

1 Introduction

Road traffic accidents constitute a major source of mortality and injury worldwide and pose substantial challenges for public health systems and sustainable development efforts. Globally, road crashes account for over 2% of total deaths (Grinerud, Aarseth, and Robertsen, 2021): this corresponds roughly to 1.19 million fatalities per year in the period 2010-2021 (Dong, Zhang, Wang, Liu, and Zhang, 2024). Beyond fatalities, the scale of non-fatal harm is considerably larger: between 20 and 50 million individuals are injured annually, many of whom suffer long-term or permanent disabilities (Schwartz, Buliung, Daniel, and Rothman, 2022). The severity of this issue is particularly pronounced in economies undergoing rapid motorization. For example, in China the expansion of the vehicle market has been accompanied by persistently high levels of traffic-related mortality (e.g., estimated at around 62,000 deaths per year Yang, Jiang, Zhou, Hitosug, and Wang (2023); Kang (2024)). Although mobility restrictions during the COVID-19 pandemic temporarily reduced accident rates, recent evidence indicates a return to pre-pandemic levels (Michelaraki, Sekadakis, Katrakazas, Ziakopoulos, and Yannis, 2023).

Understanding and mitigating this burden requires a system-level perspective on how urban road networks operate. In line with the objectives of Sustainable Development Goal 11, assessing the performance of urban traffic systems is essential for identifying inefficiencies and improving the allocation of scarce resources. Inefficiencies in these systems generate significant economic and social costs. Traffic congestion alone is estimated to reduce regional output by between 1% and 3% of GDP in many cities (Bao, Zhai, and Shen, 2022). Moreover, a substantial share of these inefficiencies can be traced back to planning and investment decisions. Inadequate infrastructure design and misallocation of resources often create persistent mismatches that are costly to correct, even frequently requiring expenditures several times larger than the initial investment (Love, Ika, Matthews, and Fang, 2020). These challenges have motivated the development of integrated evaluation frameworks capable of jointly assessing infrastructure provision, operational performance, and maintenance practices, thereby supporting more informed decision-making in transport policy and urban planning (Chen, Guo, Zhu, Wang, and He, 2022). While detailed traffic flow models provide valuable insights into micro-level dynamics and behavioral heterogeneity (Zhang, Smirnova, Bogdanova, Zhu, and Smirnov, 2018; Hu, Smirnova, Zhang, Smirnov, and Zhu, 2021), such approaches are not considered here due to data limitations.

To enable this integrated assessment, specific analytical tools are required to measure and compare the operational efficiency of different road traffic systems. The literature on productive efficiency analysis provides a robust methodological foundation for this purpose. Within this domain, nonparametric methods such as Data Envelopment Analysis (DEA) and Free Disposal Hull (FDH) have been extensively employed to evaluate transportation system performance (e.g., Cherchye, De Rock, Saelens, Verschelde, and Roets (2024), Farzipoor Saen, Karimi, and Fathi (2025), Chen, Yi, Yang, Liu, and Qu (2026)). Despite their widespread utilization, these conventional frontier estimators are known to be susceptible to overfitting, as they tend to envelop the sample data too closely, thereby potentially distorting efficiency measurements (see, e.g., Aparicio, Esteve, Rodriguez-Sala, and Zofio (2021) and Joo (2025)).

Recent methodological advancements have introduced machine learning (ML) techniques into production analysis to address these limitations. Notably, Esteve, Aparicio, Rabasa, and Rodriguez-Sala (2020) develop Efficiency Analysis Trees (EAT), a regression tree-based method that constructs production frontiers adhering to micro-economic principles while mitigating overfitting. Its convexified extension (Convexified Efficiency Analysis Trees (CEAT)) furthermore incorporates the convexity assumption. Both EAT and CEAT offer a flexible alternative to traditional estimators. Empirical applications demonstrate the value of these ML techniques across diverse contexts, including dynamic efficiency measurement (see Aparicio, Esteve, and Kapelko (2023)), wastewater treatment assessment (e.g., Maziotis, Sala-Garrido, Mocholi-Arce, and Molinos-Senante (2023)), higher education evaluation (for instance, Zofio, Aparicio, Barbero, and Zabala-Iturriagagoitia (2024)), and municipal solid waste management (for example, Sala-Garrido, Mocholi-Arce, Molinos-Senante, and Maziotis (2024)). More recently, Aparicio, Zofío, Noonan, Thompson, and Woronkowicz (2025) evaluate the efficiency of U.S. museums using a novel combination of CEAT and DEA. However, their evaluation implicitly assumes that the underlying production technologies satisfy convexity without empirically testing this assumption. Guillen, Charles, and Aparicio (2025) provide a foundational framework for integrating these ML approaches with traditional efficiency analysis, while Gaganis, Pasiouras, Tasiou, and Zopounidis (2026) introduce a novel measure of environmental revenue efficiency and apply CEAT to mitigate overfitting in DEA. Nevertheless, these studies do not examine whether the convexity assumption embedded in CEAT is consistent with the underlying data-generating process. By failing to test for convexity, they overlook the possibility that the true technology may be nonconvex, which renders CEAT misspecified and its efficiency scores potentially misleading. Despite these advances, the application of EAT and CEAT to urban road traffic systems remains unexplored.

The by-production framework –introduced by Murty, Russell, and Levkoff (2012)– offers a theoretically rigorous approach to modeling technologies that jointly generate desirable and undesirable outputs through the specification of distinct sub-technologies. While this framework has informed subsequent theoretical and empirical work, its integration with ML-based frontier estimators has yet to be systematically examined. Ray, Mukherjee, and Venkatesh (2018) build on this conceptual foundation by treating the good output as separable from the bad output and by assuming joint disposability between the bad output and the pollution-generating input, thereby departing from conventional assumptions of weak disposability and null jointness. In a related contribution, Ang, Kerstens, and Sadeghi (2023) employ Hicks-Moorsteen productivity indices to measure energy productivity change, defined as the ratio of aggregate output change to energy use change, and greenhouse gas emission intensity change, expressed as the ratio of greenhouse gas emission change to polluting input change. These same authors report convex and nonconvex productivity estimates, but do not explicitly test for convexity. Despite these theoretical and empirical developments, the geometric properties of the sub-technologies embedded in the by-production framework, particularly whether these satisfy convexity, remain untested in the context of EAT and CEAT.

This gap is consequential, as the convexity or nonconvexity of these technologies directly informs the choice of appropriate frontier estimators.

It is well-known that nonconvexities in the technology may appear for various reasons (e.g., Farrell (1959) for an early overview): there may be indivisibilities in inputs and/or outputs; increasing returns to scale and economies of scale as well as economies of specialization lead to violations of convexity; positive or negative production externalities create nonconvexities in the technology of the recipient organisation; etc. However, in the vast majority of empirical production analyses the assumption of convexity is maintained due to underlying perfect time divisibility hypothesis (see Shephard (1970, p. 15)). If perfect time divisibility is not valid and setup times and resulting setup costs are non-zero, then convexity does not hold. Thus, convexity is an empirical assumption that necessitates statistical testing to assess its validity. Unfortunately, the majority of empirical production research ignores this necessary testing step.

This paper addresses these gaps by introducing a novel methodology that integrates ML with productive efficiency analysis to measure operational inefficiency in urban road traffic systems. Grounded in the by-production framework, we decompose the production process into two distinct sub-technologies: one governing desirable outputs (transportation services) and another governing undesirable outputs (traffic accidents, casualties, and property damage). We employ EAT and CEAT to estimate the respective production frontiers based on panel data from 119 Chinese cities over the period 2014-2021.

Our empirical strategy builds on the framework proposed by Guillen, Aparicio, Kapelko, and Esteve (2025) who employ frontier methods (DEA and FDH) and ML variants (CEAT and EAT) to assess environmental efficiency and conduct Li-test-based comparisons across estimators. Our study differs from theirs in several respects. First, we apply their approach to the context of urban road traffic systems, a previously unexplored empirical setting. Second, we introduce two key extensions. First, we compare the performance of the same four estimators under alternative specifications of pollution-generating inputs, considering both the exclusion and the inclusion of street lighting. Second, we refine the methodological framework by implementing a formal statistical testing procedure that goes beyond the original analysis. In particular, following Guillen, Aparicio, Kapelko, and Esteve (2025), we also apply the nonparametric Li-test (Li, 1996; Li, Maasoumi, and Racine, 2009) to evaluate differences in the distributions of efficiency scores across estimators. Furthermore, we employ for the first time in the by-production context the convexity tests developed by Kneip, Simar, and Wilson (2016) and Simar and Wilson (2020) (henceforth KSW) to empirically examine the geometric properties of the two sub-technologies. By combining these tests with the decision rule recently proposed in Simar and Wilson (2026), we provide a novel statistical foundation for estimator selection in by-production models.

Therefore, relative to Guillen, Aparicio, Kapelko, and Esteve (2025), this paper makes three main contributions. First, it introduces EAT and CEAT methods into the operational efficiency assessment of the urban road traffic system, thereby extending the scope of ML techniques in productive efficiency analysis to a new empirical domain. Second, by subjecting both sub-technologies (associated with desirable and undesirable outputs) to convexity testing, it uncovers the inherent trade-off between "service improvement" and "negative externality control" in urban road traffic systems, while providing statistical justification for estimator selection. Third, we provide the first empirical test of convexity for each constituent sub-technology separately within the by-production framework by applying the KSW procedure. This allows us to statistically assess whether convexity is a tenable assumption for the desirable-output and undesirable-output sub-technologies, rather than imposing it a priori.

The remainder of this contribution is organized as follows. Section 2 introduces the methodological foundations, including FDH, DEA, EAT, CEAT, as well as the by-production framework. Section 3 presents the integration of these ML techniques into the by-production approach. Section 4 describes the data set and variables. Section 5 reports and discusses the empirical results, including operational inefficiency scores, distributional comparisons via Li-tests, and KSW convexity tests for both sub-technologies. Section 6 offers conclusions and outlines directions for future research.

2 Methodologies

This section opens with a review of the core concepts underlying convex and nonconvex production technologies, before introducing the by-production approach proposed by Murty, Russell, and Levkoff (2012). This structured overview establishes a foundation for the methodological comparisons that follow.

2.1 Convex and Nonconvex Production Technologies

Consider a sample $\mathcal{A}$ consisting of n Decision-Making Units (DMUs). Each DMU_i uses an input vector $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{mi}) \in \mathbb{R}^m_+$ to generate an output vector $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{si}) \in \mathbb{R}^s_+$. The underlying production possibility set or technology is defined as the set of feasible input–output combinations, namely $\boldsymbol{\Psi} = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{m+s}_+ : \boldsymbol{x} \text{ can produce at least } \boldsymbol{y}\}$.

A key distinction in nonparametric efficiency analysis lies in whether convexity is imposed on the technology. Traditional approaches such as Free Disposal Hull (FDH) and Data Envelopment Analysis (DEA) provide alternative representations of $\boldsymbol{\Psi}$ based on different structural assumptions (see Deprins, Simar, and Tulkens (1984); Banker, Charnes, and Cooper (1984)).

The FDH estimator constructs the technology set using only observed activities, relying on three basic principles. First, free disposal ensures that if $(\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{\Psi}$, then any $(\boldsymbol{x}', \boldsymbol{y}')$ satisfying $\boldsymbol{x}' \geq \boldsymbol{x}$ and $\boldsymbol{y}' \leq \boldsymbol{y}$ also belongs to $\boldsymbol{\Psi}$. Second, the envelopment condition requires that all observed DMUs are feasible. Third, the minimal extrapolation principle guarantees that the resulting technology is the smallest set satisfying these properties.

By contrast, DEA extends FDH by also incorporating the axiom of convexity. Specifically, for any $(\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}') \in \boldsymbol{\Psi}$ and any $\lambda \in [0, 1]$, their convex combination is also feasible. While this assumption yields a smoother frontier, it may impose restrictions that are not consistent with certain production environments.

Technical inefficiency is evaluated using the directional distance function (DDF) introduced by Chambers, Chung, and Färe (1998). For a given direction vector $\boldsymbol{g} = (\boldsymbol{g}_0^x, \boldsymbol{g}_0^y) \in \mathbb{R}_+^{m+s}$ and an observation $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, the DDF is defined as:

$$\vec{D}(\boldsymbol{x}_0, \boldsymbol{y}_0; \boldsymbol{g}_0^x, \boldsymbol{g}_0^y, \boldsymbol{\Psi}) = \max\{\beta_0 \in \mathbb{R} : (\boldsymbol{x}_0 - \beta_0 \boldsymbol{g}_0^x, \boldsymbol{y}_0 + \beta_0 \boldsymbol{g}_0^y) \in \boldsymbol{\Psi}\}. \tag{1}$$

When FDH is used to approximate $\boldsymbol{\Psi}$, then the corresponding binary mixed-integer optimization problem becomes:

$$\begin{aligned}
&\vec{D}^{FDH}(\boldsymbol{x}_0, \boldsymbol{y}_0; \boldsymbol{g}_0^x, \boldsymbol{g}_0^y) = \max \beta_0 \\
s.t. \; &\sum_{i=1}^{n} \lambda_i x_{ji} \leq x_{j0} - \beta_0 g_{j0}^x, \quad j = 1, \ldots, m \\
&\sum_{i=1}^{n} \lambda_i y_{ri} \geq y_{r0} + \beta_0 g_{r0}^y, \quad r = 1, \ldots, s \\
&\sum_{i=1}^{n} \lambda_i = 1 \\
&\lambda_i \in \{0, 1\}, \quad i = 1, \ldots, n
\end{aligned} \tag{2}$$

In the DEA formulation, the structure of model (2) remains unchanged except that the binary constraints are relaxed to $\lambda_i \geq 0$. In both cases, the theoretical technology $\boldsymbol{\Psi}$ in model (1) is replaced by its empirical counterparts, namely $\hat{\boldsymbol{\Psi}}^{\text{FDH}}$ and $\hat{\boldsymbol{\Psi}}^{\text{DEA}}$.

Beyond these classical approaches, recent developments have introduced machine learning techniques to represent production technologies. In particular, Esteve, Aparicio, Rabasa, and Rodriguez-Sala (2020) propose EAT, which construct a piecewise representation of the frontier through recursive partitioning of the input space. This approach preserves free disposability and envelopment while relaxing the minimal extrapolation principle, thereby reducing the overfitting issues often associated with FDH (see Esteve, Aparicio, Rabasa, and Rodriguez-Sala (2020); Tsionas (2022)). The convexified extension CEAT furthermore incorporates the convexity assumption in addition to the free disposability and envelopment axioms, thereby relaxing the minimal extrapolation principle to reduce overfitting associated with DEA.

The EAT algorithm proceeds by iteratively splitting the sample. For a given node t, the method selects an input variable j and a threshold $s_j \in S_j$ to partition the data into two subsets, minimizing the sum of mean squared errors across child nodes. The process continues until a stopping rule is met, typically when the node size satisfies $n(t) \leq 5$ (Breiman, Friedman, Olshen, and Stone, 2017).

Each terminal node defines a region in the input space, referred to as its support: $supp(t) = \{\boldsymbol{x} \in \mathbb{R}_+^m : a_j^t \leq x_j < b_j^t, j = \{1, \ldots, m\}\}$. Based on these regions, a Pareto-dominance relation between nodes can be established, which determines the structure of the efficient frontier.

After growing a maximal tree $T_{max}(\mathcal{A})$, a pruning procedure based on the error-complexity criterion is applied to obtain a reduced tree $T^*(\mathcal{A})$. Let $\tilde{T^*}(\mathcal{A})$ denote the set of terminal nodes of the pruned tree. The associated multidimensional predictor is defined as $\boldsymbol{d}_{T^*(\mathcal{A})}(\boldsymbol{x})$, with components given by $\boldsymbol{d}_{rT^*(\mathcal{A})}(\boldsymbol{x}) = \sum_{t\in\tilde{T^*}(\mathcal{A})} y_r(t)I(\boldsymbol{x} \in supp(t))$.

Using this structure, the DDF under EAT can be computed through the following binary mixed-integer linear program:

$$\begin{aligned}
& \vec{D}^{EAT}(\boldsymbol{x}_0, \boldsymbol{y}_0; \boldsymbol{g}_0^x, \boldsymbol{g}_0^y) = \max \beta_0 \\
s.t. \quad & \sum_{t\in\tilde{T^*}(\mathcal{A})} \lambda_t a_j^t \leq x_{j0} - \beta_0 g_{j0}^x, \quad j = 1, \ldots, m \\
& \sum_{t\in\tilde{T^*}(\mathcal{A})} \lambda_t d_{rT^*}(\boldsymbol{a}^t) \geq y_{r0} + \beta_0 g_{r0}^y, \quad r = 1, \ldots, s \\
& \sum_{t\in\tilde{T^*}(\mathcal{A})} \lambda_t = 1 \\
& \lambda_t \in \{0, 1\}, t \in \tilde{T^*}(\mathcal{A})
\end{aligned} \tag{3}$$

Unlike model (2), which is based on observed data, model (3) relies on a constructed dataset $\hat{\mathcal{A}} = \{(\boldsymbol{a}^t, \boldsymbol{d}_{T^*}(\boldsymbol{a}^t))\}_{t\in\tilde{T^*}(\mathcal{A})}$ derived from the tree structure.

A convexified version of EAT, referred to as CEAT, is obtained by relaxing the binary constraints to $\lambda_t \geq 0$ (Aparicio, Esteve, Rodriguez-Sala, and Zofio, 2021). This extension bridges the gap between nonconvex and convex representations by combining the flexibility of tree-based partitioning with the smoothing effect of convexity.

In both EAT and CEAT formulations, the DDF is computed by replacing $\boldsymbol{\Psi}$ in model (1) with the corresponding estimated technology.

2.2 By-production Technology

The by-production framework proposed by Murty, Russell, and Levkoff (2012) accounts for the joint generation of desirable and undesirable outputs by recognizing that these outputs may arise from distinct underlying mechanisms. Instead of representing production through a single technology, this approach decomposes the overall production set into two interrelated sub-technologies. One sub-technology, denoted $\boldsymbol{\Psi}_1$, characterizes the generation of desirable outputs, while the other, $\boldsymbol{\Psi}_2$, describes the formation of undesirable outputs. The global production technology is then defined as their intersection, $\boldsymbol{\Psi} = \boldsymbol{\Psi}_1 \cap \boldsymbol{\Psi}_2$, ensuring that feasible production plans satisfy both processes simultaneously.

Consider n observed DMUs indexed by $i = \{1, \ldots, n\}$. For a given unit, let $\boldsymbol{x}_0 = (x_{10}, \ldots, x_{m0}) \in \mathbb{R}_+^m$ denote the input vector, $\boldsymbol{y}_0 = (y_{10}, \ldots, y_{s0}) \in \mathbb{R}_+^s$ the desirable outputs, and $\boldsymbol{z}_0 = (z_{10}, \ldots, z_{s'0}) \in \mathbb{R}_+^{s'}$ the undesirable outputs.

Following Murty, Russell, and Levkoff (2012), inputs are partitioned into non-polluting inputs $\boldsymbol{q} \in \mathbb{R}_+^{m_1}$ and pollution-generating inputs $\boldsymbol{p} \in \mathbb{R}_+^{m_2}$, with $m_1 + m_2 = m$. The former typically includes conventional production factors such as labor and capital, whereas the latter captures inputs associated with the generation of undesirable by-products, such as energy inputs linked to

CO_2 emissions. Within this setting, the two sub-technologies are defined as:

$$\boldsymbol{\Psi}_1 = \{(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}_+^{m_1+m_2+s+s'} : h(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{y}) \leq 0\}, \tag{4}$$

$$\boldsymbol{\Psi}_2 = \{(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}_+^{m_1+m_2+s+s'} : \boldsymbol{z} \geq l(\boldsymbol{p})\}. \tag{5}$$

where h and l are continuously differentiable functions.

Under the FDH framework, these sub-technologies are approximated empirically using observed data as follows:

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_1^{FDH} = \{(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}_+^{m_1+m_2+s+s'} : \sum_{i=1}^{n} \lambda_i \boldsymbol{q}_i \leq \boldsymbol{q}, \sum_{i=1}^{n} \lambda_i \boldsymbol{p}_i \leq \boldsymbol{p}, \sum_{i=1}^{n} \lambda_i \boldsymbol{y}_i \geq \boldsymbol{y}, \\ \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \in \{0,1\}, \forall i\}\end{aligned} \tag{6}$$

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_2^{FDH} = \{(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}_+^{m_1+m_2+s+s'} : \sum_{i=1}^{n} \mu_i \boldsymbol{p}_i \geq \boldsymbol{p}, \sum_{i=1}^{n} \mu_i \boldsymbol{z}_i \leq \boldsymbol{z}, \\ \sum_{i=1}^{n} \mu_i = 1, \mu_i \in \{0,1\}, \forall i\},\end{aligned} \tag{7}$$

where λ_i and μ_i are binary intensity variables.

To evaluate inefficiency relative to each sub-technology, we employ the output-oriented directional distance function proposed by Murty, Russell, and Levkoff (2012). Combined with nonparametric estimators such as FDH, this approach enables the empirical implementation of the DDF (see Aparicio, Kapelko, and Zofio (2020); Aparicio, Esteve, Rodriguez-Sala, and Zofio (2021)). Specifically, the DDF for the good-output and bad-output sub-technologies can be obtained from the following two binary mixed-integer linear programs:

$$\begin{aligned} & \vec{B}_{\Psi_1}^{FDH}(\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{y}_0; \hat{\boldsymbol{\Psi}}_1^{FDH}) = Max\, \beta^{\Psi_1} \\ s.t.\ & \sum_{i=1}^{n} \lambda_i q_{ji} \leq q_{j0}, j = 1, \ldots, m_1 \\ & \sum_{i=1}^{n} \lambda_i p_{ji} \leq p_{j0}, j = m_1 + 1, \ldots, m \\ & \sum_{i=1}^{n} \lambda_i y_{ri} \geq y_{r0} + \beta^{\Psi_1} y_{r0}, r = 1, \ldots, s \\ & \beta^{\Psi_1} \geq 0 \\ & \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \in \{0,1\}, i = 1, \ldots, n, \end{aligned} \tag{8}$$

and

$$
\begin{aligned}
& \vec{B}^{FDH}_{\Psi_2}(\boldsymbol{p}_0, \boldsymbol{z}_0; \hat{\boldsymbol{\Psi}}^{FDH}_2) = Max\, \beta^{\Psi_2} \\
s.t.\ & \sum_{i=1}^{n} \mu_i p_{ji} \geq p_{j0}, j = m_1 + 1, \ldots, m \\
& \sum_{i=1}^{n} \mu_i z_{ki} \leq z_{k0} - \beta^{\Psi_2} z_{k0}, k = 1, \ldots, s' \\
& \beta^{\Psi_2} \geq 0 \\
& \sum_{i=1}^{n} \mu_i = 1, \mu_i \in \{0, 1\}, i = 1, \ldots, n.
\end{aligned}
\tag{9}
$$

Model (8) evaluates the potential expansion of desirable outputs while holding inputs and undesirable outputs fixed, whereas model (9) captures the feasible contraction of undesirable outputs. In both cases, the use of binary intensity variables implies that only observed production plans are used to construct the reference frontier.

The resulting inefficiency measures satisfy $\beta^{\Psi_1 *} \in [0, \infty)$ and $\beta^{\Psi_2 *} \in [0, 1]$. Combining both components, the overall DDF under the by-production framework is given by $\vec{B}^{FDH}(\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{y}_0, \boldsymbol{z}_0) = \delta^{\Psi_1}(1 - \frac{1}{1+\beta^{\Psi_1 *}}) + \delta^{\Psi_2}\beta^{\Psi_2 *}$, where δ^{Ψ_1} and δ^{Ψ_2} are non-negative weights satisfying $\delta^{\Psi_1} + \delta^{\Psi_2} = 1$. These parameters reflect the relative emphasis placed on expanding desirable outputs versus reducing undesirable outputs.

The DEA-based counterpart is obtained by replacing the binary constraints in models (8) and (9) with $\lambda_i \geq 0$ and $\mu_i \geq 0$, respectively. The corresponding overall DDF, denoted $\vec{B}^{DEA}$, is defined analogously using the optimal values from the DEA formulations.

We do not assume a priori the geometric properties of the two sub-technologies, in particular whether these satisfy convexity. Following recent advances in nonparametric efficiency analysis (Kneip, Simar, and Wilson, 2016; Simar and Wilson, 2020), we treat convexity as an empirical hypothesis to be tested separately for $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$. This distinction is particularly important in the by-production context, where the desirable-output and undesirable-output sub-technologies arise from distinct physical and economic mechanisms. Imposing a common convexity assumption on both sub-technologies without empirical testing risks misspecifying one or both frontiers: this may misguide policy prioritisation.

3 Solving the By-production Technology through ML Techniques

Following closely Guillen, Aparicio, Kapelko, and Esteve (2025, Section 3), this section develops a machine-learning implementation of the by-production framework by combining tree-based frontier estimation with the two sub-technologies introduced above. The objective is to estimate production structures in the presence of both desirable and undesirable outputs while preserving the interpretation of the by-production model. Regression trees provide a flexible device for this purpose because they partition the input space recursively and can accommodate nonlinearities, heterogeneity, and complex interactions across production factors. In what follows, the two sub-technologies are estimated separately: the first one describes the production of desirable outputs, whereas the second one captures the generation of undesirable outputs.

Our strategy relies on the EAT algorithm to build an initial tree representation of each sub-technology and on the associated pruning step to avoid overfitting. Once the final trees are obtained, the resulting piecewise frontier can be used directly under a nonconvex specification. When convexity is required, the CEAT formulation is adopted by relaxing the binary intensity constraints. In this way, the same tree-based architecture supports both nonconvex and convex versions of the by-production model.

We begin with the good-output sub-technology. Since the original EAT estimator is based on a stepwise frontier, model (8) must first be translated into the input-output language required by the tree algorithm. This can be done by inspecting the direction of the inequalities in model (8). The first two sets of constraints have the standard input orientation, whereas the third one has an output orientation. Accordingly, in the tree-building stage, $\boldsymbol{q}$ and $\boldsymbol{p}$ are handled as inputs, while $\boldsymbol{y}$ is treated as the output vector.

Running the EAT procedure on this specification yields the optimal tree $T_1^*(\mathcal{A})$, its set of terminal nodes, the corresponding supports, and the predictor $\boldsymbol{d}_{T_1^*(\mathcal{A})}(\boldsymbol{x})$. For each terminal node t, the support is described by the lower and upper bounds $\boldsymbol{a}^t_{T_1^*(\mathcal{A})}$ and $\boldsymbol{b}^t_{T_1^*(\mathcal{A})}$. The collection $\{\boldsymbol{a}^t_{T_1^*(\mathcal{A})}\}_{t\in\tilde{T}_1^*(\mathcal{A})}$ provides the input coordinates of the virtual points $\{(\boldsymbol{a}^t_{T_1^*(\mathcal{A})}, \boldsymbol{d}_{T_1^*(\mathcal{A})}(\boldsymbol{a}^t_{T_1^*(\mathcal{A})}))\}_{t\in\tilde{T}_1^*(\mathcal{A})}$, which replace the original sample points when computing the directional distance function. Therefore, the nonconvex by-production formulation for the desirable-output technology is written as:

$$
\begin{aligned}
& \vec{B}^{EAT}_{\Psi_1}(\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{y}_0; \hat{\boldsymbol{\Psi}}_1^{EAT}) = Max\, \beta^{\Psi_1} \\
s.t. \quad & \sum_{t\in\tilde{T}_1^*(\mathcal{A})} \lambda_t a^t_{jT_1^*(\mathcal{A})} \le q_{j0}, j = 1,\dots,m_1 \\
& \sum_{t\in\tilde{T}_1^*(\mathcal{A})} \lambda_t a^t_{jT_1^*(\mathcal{A})} \le p_{j0}, j = m_1+1,\dots,m \\
& \sum_{t\in\tilde{T}_1^*(\mathcal{A})} \lambda_t d_{rT_1^*(\mathcal{A})}(\boldsymbol{a}^t_{T_1^*(\mathcal{A})}) \ge y_{r0} + \beta^{\Psi_1} y_{r0}, r = 1,\dots,s \\
& \beta^{\Psi_1} \ge 0 \\
& \sum_{t\in\tilde{T}_1^*(\mathcal{A})} \lambda_t = 1, \lambda_t \in \{0,1\}, t \in \tilde{T}_1^*(\mathcal{A}).
\end{aligned}
\tag{10}
$$

Relative to model (8), the right-hand side of model (10) remains unchanged, whereas the left-hand side of model (10) is reconstructed from the virtual observations induced by the terminal nodes of $T_1^*(\mathcal{A})$. Hence, feasible comparisons are no longer formed from original DMUs but from the lower bounds of terminal-node supports together with the predicted output values attached to those supports.

We next turn to the bad-output sub-technology. In this case, model (9) must again be rewritten in a form compatible with the EAT classification scheme. The first constraint has an output-type role, while the second one behaves as an input-type restriction. Consequently, the variables $\boldsymbol{p}$ are treated as outputs and $\boldsymbol{z}$ as inputs for the purpose of tree estimation. Applying the EAT algorithm under this arrangement produces a second optimal tree, denoted $T_2^*(\mathcal{A})$, together with its terminal nodes, supports, and predictor. The corresponding nonconvex formulation for the undesirable-output technology is then expressed as shown below:

$$
\begin{aligned}
& \vec{B}_{\Psi_2}^{EAT}(\boldsymbol{p}_0, \boldsymbol{z}_0; \hat{\boldsymbol{\Psi}}_2^{EAT}) = Max\, \beta^{\Psi_2} \\
s.t. \quad & \sum_{t \in \tilde{T}_2^*(\mathcal{A})} \mu_t d_{jT_2^*(\mathcal{A})}(\boldsymbol{a}_{T_2^*(\mathcal{A})}^t) \geq p_{j0}, j = m_1 + 1, \ldots, m \\
& \sum_{t \in \tilde{T}_2^*(\mathcal{A})} \mu_t a_{kT_2^*(\mathcal{A})}^t \leq z_{k0} - \beta^{\Psi_2} z_{k0}, k = 1, \ldots, s' \\
& \beta^{\Psi_2} \geq 0 \\
& \sum_{t \in \tilde{T}_2^*(\mathcal{A})} \mu_t = 1, \mu_t \in \{0, 1\}, t \in \tilde{T}_2^*(\mathcal{A}).
\end{aligned}
\tag{11}
$$

Again, the reference set in model (11) is formed by virtual observations rather than by the original sample. More precisely, the lower bounds of the terminal-node supports in $\tilde{T}_2^*(\mathcal{A})$ act as proxies for undesirable outputs, while the associated predicted values replace the pollution-generating inputs on the left-hand side of the program.

Once models (10) and (11) have been solved, the global by-production inefficiency score under EAT is obtained by combining the two optimal values according to $\vec{B}^{EAT}(\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{y}_0, \boldsymbol{z}_0) = \delta^{\Psi_1}(1 - \frac{1}{1+\beta^{\Psi_1 *}}) + \delta^{\Psi_2}\beta^{\Psi_2 *}$.

The convexified version follows immediately. To move from EAT to CEAT, it is enough to replace the binary intensity variables in models (10) and (11) with nonnegative intensity variables. In model (10), this means substituting $\lambda_t \in \{0, 1\}$ with $\lambda_t \geq 0$, whereas in model (11) the condition $\mu_t \in \{0, 1\}$ is replaced by $\mu_t \geq 0$. The resulting CEAT formulations for the two sub-technologies are shown as follows:

$$
\begin{aligned}
& \vec{B}_{\Psi_1}^{CEAT}(\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{y}_0; \hat{\boldsymbol{\Psi}}_1^{CEAT}) = Max\, \beta^{\Psi_1} \\
s.t. & \sum_{t \in \tilde{T}_1^*(\mathcal{A})} \lambda_t a_{jT_1^*(\mathcal{A})}^t \leq q_{j0}, j = 1, \ldots, m_1 \\
& \sum_{t \in \tilde{T}_1^*(\mathcal{A})} \lambda_t a_{jT_1^*(\mathcal{A})}^t \leq p_{j0}, j = m_1 + 1, \ldots, m \\
& \sum_{t \in \tilde{T}_1^*(\mathcal{A})} \lambda_t d_{rT_1^*(\mathcal{A})}(\boldsymbol{a}_{T_1^*(\mathcal{A})}^t) \geq y_{r0} + \beta^{\Psi_1} y_{r0}, r = 1, \ldots, s \\
& \beta^{\Psi_1} \geq 0 \\
& \sum_{t \in \tilde{T}_1^*(\mathcal{A})} \lambda_t = 1, \lambda_t \geq 0, t \in \tilde{T}_1^*(\mathcal{A}),
\end{aligned}
\tag{12}
$$

and

$$
\begin{aligned}
& \vec{B}_{\Psi_2}^{CEAT}(\boldsymbol{p}_0, \boldsymbol{z}_0; \hat{\boldsymbol{\Psi}}_2^{CEAT}) = Max\, \beta^{\Psi_2} \\
s.t. & \sum_{t \in \tilde{T}_2^*(\mathcal{A})} \mu_t d_{jT_2^*(\mathcal{A})}(\boldsymbol{a}_{T_2^*(\mathcal{A})}^t) \geq p_{j0}, j = m_1 + 1, \ldots, m \\
& \sum_{t \in \tilde{T}_2^*(\mathcal{A})} \mu_t a_{kT_2^*(\mathcal{A})}^t \leq z_{k0} - \beta^{\Psi_2} z_{k0}, k = 1, \ldots, s' \\
& \beta^{\Psi_2} \geq 0 \\
& \sum_{t \in \tilde{T}_2^*(\mathcal{A})} \mu_t = 1, \mu_t \geq 0, t \in \tilde{T}_2^*(\mathcal{A}).
\end{aligned}
\tag{13}
$$

Finally, $\vec{B}^{CEAT}(\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{y}_0, \boldsymbol{z}_0) = \delta^{\Psi_1}(1 - \frac{1}{1+\beta^{\Psi_1 *}}) + \delta^{\Psi_2}\delta^{\Psi_2 *}$.

4 Description of the Secondary Sample

In this section, we illustrate the practical applicability of our developed methods for measuring operational inefficiency by applying them to the operational performance of urban road traffic systems in 119 cities from 2014 to 2021 earlier studied in Chen, Yi, Yang, Liu, and Qu (2026). Socio-economic data on urban road traffic systems are obtained from the *China Urban Construction Statistical Yearbook* and the *China City Statistical Yearbook* (2015-2022). Data on traffic accidents, including the number of accidents, casualties, and direct property damage, are primarily drawn from the *China Statistical Yearbook of Road Traffic Accidents* (2015-2016). However, as comprehensive disclosure of traffic accident data in official statistical reports has been limited since 2016, we address this data gap by reviewing statistical yearbooks for approximately 300 prefecture-level cities across China from 2015 to 2022. Our review reveals that some cities continue to publish traffic accident data on an annual basis. In addition, missing data are supplemented with information obtained from local transportation bureaus, prefecture-level city statistical bureaus, and regional economic development yearbooks. Through this process, we construct a unique dataset of traffic accidents tailored to the needs of this study. The final balanced sample comprises 119 cities observed over the period 2014-2021.

In evaluating the operational efficiency of urban road traffic systems within the by-production framework introduced in Section 2.2, we categorize inputs based on their role in the production process and their association with undesirable outputs. Non-polluting inputs are essential for infrastructure provision and maintenance without directly generating environmental externalities. These primarily include Land for Road Transportation Facilities (unit: square kilometer), Expenditure for Road Transportation (unit: 10,000 yuan), Road Length (unit: kilometer), and, in most cases, Number of Street Lights (unit: piece).

Moreover, the primary pollution-generating input in our context is the number of vehicles owned by civilians, since vehicle operation is directly associated with traffic congestion and safety risks. The classification of street lights is less clear-cut. In most analytical settings, street lights are treated as a non-polluting input that enhances road infrastructure and safety. However, in this study we also consider a specification that includes street lights as a second pollution-generating input. This choice is motivated by empirical evidence and the methodological requirements of the by-production framework. As shown in the correlation analysis reported in Appendix A, the number of civilian-owned vehicles exhibits the strongest and most consistent correlation with undesirable outputs across all three statistical tests, validating it as the most robust candidate for a pollution-generating input. The number of street lights emerges as the second-most correlated variable. Because the by-production framework requires pollution-generating inputs to be non-negatively correlated with undesirable outputs along the frontier, including street lights, while not strictly necessary, provides a useful robustness check. Accordingly, depending on the analytical context, we consider two specifications: (i) one with a single pollution-generating input (vehicles only), and (ii) another with two such inputs (vehicles and street lights).

Consistent with the by-production framework, we further distinguish between desirable (good) outputs and undesirable (bad) outputs. The desirable outputs are total road passenger volume (unit: 10,000 person) and freight volume (unit: 10,000 ton), reflecting the primary transportation services provided by the urban road system. The undesirable outputs include three indicators of negative externalities: (i) number of traffic accidents (unit: piece), (ii) number of fatalities and injuries (unit: person), and (iii) direct property damage caused by traffic accidents (unit: 10,000 yuan). This distinction enables us to model the joint production of transportation services and negative externalities within a unified analytical framework.

Given the large number of cities, Table 1 exhibits detailed descriptive statistics of inputs-outputs quantities for each year, as well as over the whole 2014-2021 period. Table 1 highlights distinct evolutionary patterns among input categories over the period 2014-2021. Non-polluting inputs exhibit consistent expansionary trends. Specifically, land allocated for road transportation infrastructure increased by 44.0 percent, from an average of 28.05 square kilometers in 2014 to 40.38 square kilometers in 2021. Likewise, total road length expanded from 13,655.80 kilometers to 16,337.82 kilometers, while the number of street lights surged by 78.7 percent, rising from 75,236.52 to 134,375.09 units. These developments reflect ongoing urbanization and systematic improvements in road network quality. In contrast, transportation expenditure displayed a fluctuating trajectory, peaking at 3.43 billion yuan in 2015 before declining to 2.44 billion yuan by 2021, a pattern that may indicate enhanced investment efficiency or fiscal restructuring. Meanwhile, civilian vehicle ownership, a key pollution-generating input, demonstrated robust growth, increasing by 92.7 percent from 601,482.92 vehicles in 2014 to 1,159,202.24 vehicles in 2021, corresponding to an average annual growth rate of approximately 9.8 percent. This rapid motorization trend underscores the rising demand for urban mobility while highlighting growing concerns over negative externalities such as traffic congestion and environmental degradation.

Table 1 also reveals a divergence between desirable and undesirable outputs over the study period. Among desirable outputs, total road freight volume remained relatively stable, fluctuating between 129.50 million and 156.60 million tons, reflecting the inelastic nature of urban logistics demand. In contrast, total road passenger volume experienced a sharp decline, falling from 72.61 million person-times in 2014 to 23.50 million person-times in 2021, a reduction of 67.6 percent. This downturn likely reflects growing competition from alternative transport modes such as high-speed rail and civil aviation, as well as the disruptive effects of the COVID-19 pandemic on travel behavior. Furthermore, undesirable outputs, by comparison, followed an upward trajectory. The number of traffic accidents rose by 52.7 percent, from 877.66 incidents in 2014 to 1,340.45 incidents in 2021. Casualties increased from 1,168.21 persons to 1,313.24 persons over the same period, while direct property damage remained relatively stable, ranging between 3.80 million and 4.56 million yuan. These trends align closely with the rapid expansion in civilian vehicle ownership, underscoring the growing traffic safety challenges that accompany accelerated motorization.

Standard deviation statistics reveal substantial city heterogeneity across all indicators. In 2021, civilian vehicle ownership exhibits a standard deviation of approximately 1.25 million vehicles, equivalent to 1.08 times its mean, underscoring pronounced disparities in motorization levels. Transportation expenditure shows even greater variation, with a standard deviation of 10.40 billion yuan in 2015, roughly three times the mean, reflecting significant differences in municipal fiscal capacity. More critically, the data unveil an efficiency paradox: despite sustained infrastructure expansion (road length +19.6%, street lights +78.7%), desirable outputs such as passenger volume decline by 67.6%, while undesirable outputs including traffic accidents rose by 52.7%. This configuration of high input, low output, and rising externalities highlights systemic operational inefficiencies and justifies the adoption of a by-production framework in this study. Collectively, these patterns suggest that Chinese urban road traffic systems are shifting from scale expansion toward efficiency enhancement, with the dual imperative of improving service delivery and mitigating negative externalities emerging as central to governance.

Table 1: Descriptive statistics of the inputs and outputs of urban road traffic system

Period	Statistics	**Inputs**					**Desirable outputs**		**Undesirable outputs**		
		Land for Road Transportation Facilities	Expenditure for Transportation	Road Length	Number of Street Lights	Number of vehicles owned by civilians	Total road passenger volume	Total road freight volume	Number of traffic accidents	Number of fatalities and injuries	Direct property damage caused by traffic accidents
2014	Average	28.05	235591.45	13655.80	75236.52	601482.92	7260.89	13273.11	877.66	1168.21	380.31
	Stand. Dev	52.15	427668.45	12305.44	89260.67	777392.90	10243.88	12328.32	972.72	1307.90	581.84
2015	Average	29.77	343403.78	14037.67	80419.74	672854.60	7940.07	15650.82	820.79	1093.81	363.07
	Stand. Dev	52.95	1040093.58	13382.67	94994.93	830109.77	17160.86	27903.62	910.78	1214.22	523.91
2016	Average	30.21	284542.14	14351.06	88173.24	752271.06	7234.53	15589.66	831.68	1092.27	398.57
	Stand. Dev	44.43	668005.94	13555.65	103952.40	897241.25	16389.49	28485.85	951.82	1231.71	633.24
2017	Average	32.23	269888.11	14591.94	88760.76	838757.45	6322.55	14334.86	829.47	1052.86	422.01
	Stand. Dev	46.60	670087.84	13993.89	104586.42	965325.18	12037.65	13339.44	948.87	1158.64	713.47
2018	Average	33.89	275785.19	14832.39	96883.80	967730.45	4918.53	15229.02	1020.03	1158.71	456.77
	Stand. Dev	48.01	645840.06	14837.10	120904.68	1032717.45	6691.57	14391.08	1447.04	1185.07	741.30
2019	Average	36.69	271393.71	15278.36	101809.19	1033669.71	4744.12	15260.39	1088.70	1152.59	416.82
	Stand. Dev	52.25	593991.31	16270.32	128614.38	1078789.56	6838.92	13729.96	2025.72	1196.04	647.01
2020	Average	38.10	282731.93	16190.89	125714.83	1088481.52	2738.72	12950.74	1087.11	1170.39	395.19
	Stand. Dev	54.54	564263.46	16971.97	231188.76	1170578.78	4804.08	12408.78	2103.61	1273.40	655.89
2021	Average	40.38	243617.19	16337.82	134375.09	1159202.24	2350.24	14477.44	1340.45	1313.24	439.22
	Stand. Dev	55.94	478856.55	17246.25	244299.88	1248293.33	6465.05	14340.82	2870.55	1425.93	719.73
2014-2021	Average	33.66	275869.19	14909.49	98921.65	889306.24	5438.71	14595.75	986.99	1150.26	408.99
	Stand. Dev	50.98	657589.21	14890.93	151983.41	1025893.58	11143.60	18244.98	1677.60	1249.27	653.91

5 Empirical Results

Our empirical strategy applies essentially a static frontier approach. More specifically, we perform independent annual frontier analyses to measure the operational efficiency of cities on a yearly basis, while we abstract from any eventual technological change of the frontiers over time. The analysis proceeds in three steps: (i) we report operational inefficiency values derived from by-production FDH and EAT models, both of which relax convexity; (ii) we examine inefficiency scores obtained from convex by-production DEA and CEAT estimators; and (iii) we present results of the Li-test and the KSW convexity test applied to the by-production DDF. Prior to presenting the inefficiency estimates, we note that our procedure involves two stages: we compute directional distance function values for the two sub-technologies using four alternative estimators, and then we subject each sub-technology to formal convexity testing via the KSW procedure. This two-stage approach allows us to determine on statistical grounds whether convexity is a suitable assumption for each sub-technology in our sample, and hence whether the convex or nonconvex estimates are more reliable for policy analysis.

5.1 Operational Inefficiency Values for By-production FDH and EAT

We begin by examining the temporal evolution of differences in operational inefficiency. Table 2 reports the mean DDF scores derived from the by-production FDH and EAT models for the 119 cities. These scores are presented horizontally annually and aggregated over the period 2014-2021 (last line) under two cases (one pollution-generating input and two pollution-generating inputs), and vertically disaggregated by the good-output technology $\mathbf{\Psi}_1$, the bad-output technology $\mathbf{\Psi}_2$, and finally the global technology $\mathbf{\Psi}$ (with $\delta^{\Psi_1} = \delta^{\Psi_2} = 0.5$).

Table 2: Mean inefficiency values from by-production FDH and EAT for $\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$, and $\mathbf{\Psi}$

Case	Period	Ψ_1^{FDH}	Ψ_2^{FDH}	Ψ^{FDH}	Ψ_1^{EAT}	Ψ_2^{EAT}	Ψ^{EAT}
One pollution-generating input	2014	0.1404	0.5483	0.3172	0.5933	0.5623	0.4132
	2015	4.6215	0.5423	0.4132	5.0787	0.5531	0.5104
	2016	4.0601	0.4636	0.3613	4.8696	0.4856	0.4749
	2017	0.1309	0.4493	0.2629	0.9877	0.5000	0.4286
	2018	0.0869	0.5026	0.2781	0.5302	0.5398	0.3979
	2019	0.1979	0.5019	0.3081	0.6365	0.5336	0.4088
	2020	0.2311	0.5219	0.3171	0.9708	0.5940	0.4714
	2021	0.2051	0.5280	0.3138	1.1913	0.5742	0.4771
	2014-2021	1.2092	0.5072	0.3215	1.8561	0.5428	0.4478
Two pollution-generating inputs	2014	0.1404	0.4941	0.2901	0.5933	0.5284	0.3963
	2015	4.6215	0.4851	0.3845	5.0787	0.5096	0.4887
	2016	4.0601	0.4207	0.3399	4.8287	0.4609	0.4625
	2017	0.1309	0.4006	0.2386	0.9877	0.4777	0.4175
	2018	0.0869	0.4191	0.2364	0.5302	0.5159	0.3859
	2019	0.1979	0.4516	0.2829	0.6365	0.5287	0.4064
	2020	0.2311	0.4465	0.2794	0.9708	0.5418	0.4453
	2021	0.2051	0.4718	0.2857	1.1913	0.5360	0.4581
	2014-2021	1.2092	0.4487	0.2922	1.8522	0.5124	0.4326

A striking feature of the FDH and EAT results is the emergence of pronounced inefficiency spikes for $\boldsymbol{\Psi}_1$ during 2015 and 2016. Under the single pollution-generating input specification, Ψ_1^{FDH} peaks at 4.6215 and 4.0601 in these years, while Ψ_1^{EAT} reaches even higher values of 5.0787 and 4.8696. This corroboration across distinct methodological families (nonconvex FDH and partition-based EAT) reinforces the interpretation that these years constitute genuine outliers in the production technology, likely driven by unique operational or policy shocks. Excluding these anomalies, the FDH estimates for $\boldsymbol{\Psi}_1$ stabilize at remarkably low levels, ranging from 0.0869 to 0.2311, suggesting that under a purely nonconvex frontier, most urban road traffic systems operate in close proximity to best practices in desirable output generation.

The contrast between FDH and EAT yields critical methodological insights regarding the relative magnitude of inefficiency across sub-technologies. For FDH, the mean inefficiency scores for $\boldsymbol{\Psi}_2$ consistently dominate those for $\boldsymbol{\Psi}_1$ in normal years. Under the single pollution-generating input case, Ψ_2^{FDH} fluctuates narrowly between 0.4493 and 0.5483, substantially exceeding the corresponding Ψ_1^{FDH} values. This pattern aligns closely with the DEA findings reported below, indicating greater potential for mitigating undesirable outputs than for expanding desirable ones when the production frontier is constructed under nonconvex or conventional convex assumptions. Conversely, the EAT estimates reverse this hierarchy: Ψ_1^{EAT} (ranging from 0.5302 to 1.1913, excluding outliers) consistently surpasses Ψ_2^{EAT} (0.4856 to 0.5940) across all years. This reversal underscores the sensitivity of efficiency rankings to the underlying frontier estimation technique. By constructing partitions via regression trees and subsequently evaluating performance within these leaves, EAT effectively re-calibrates the benchmark for desirable output production, thereby revealing latent inefficiencies that remain concealed under the less structured FDH frontier.

Temporally, the FDH and EAT estimators display divergent dynamics. The Ψ_1^{FDH} series, once purged of the 2015-2016 anomalies, exhibits near stationarity with no discernible monotonic trend, mirroring the stability observed in the DEA estimates below. By contrast, Ψ_1^{EAT} captures more pronounced intertemporal variation: after the post-2016 decline from the outlier peaks, it reaches a trough in 2018 (0.5302) before ascending steadily to 1.1913 by 2021. This U-shaped pattern, albeit less amplified than in the CEAT results, suggests that the tree-based partitioning mechanism inherent in EAT is more adept at tracking medium-term efficiency fluctuations than the purely nonconvex FDH. The $\boldsymbol{\Psi}_2$ estimates for both methods remain remarkably stable over time, reinforcing the conclusion that the efficiency frontier for undesirable output abatement is less susceptible to temporal shocks and methodological perturbations.

Regarding the specification of two pollution-generating inputs, the patterns remain broadly consistent with prior findings. The inclusion of street lights as an additional pollution-generating input yields marginally lower mean inefficiency scores for $\boldsymbol{\Psi}_2$ under both FDH and EAT, while leaving $\boldsymbol{\Psi}_1$ almost unaffected. For instance, the full-period mean for Ψ_2^{FDH} declines from 0.5072 to 0.4487, and for Ψ_2^{EAT} from 0.5428 to 0.5124, when moving from the single to the two-input specification. This adjustment reflects a re-calibration of the benchmark against which bad output performance is assessed, yet the overarching temporal trends and the relative ordering between $\boldsymbol{\Psi}_1$

and $\boldsymbol{\Psi}_2$ within each estimator remain robust to this alteration.

5.2 Operational Inefficiency Values for By-production DEA and CEAT

We now turn to the results obtained from the by-production DEA and CEAT estimators, which incorporate convexity. Table 3 presents the mean DDF values for $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$, and $\boldsymbol{\Psi}$ under the same two pollution-generating input specifications.

Table 3: Mean inefficiency values from by-production DEA and CEAT for $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$, and $\boldsymbol{\Psi}$

Case	Period	Ψ_1^{DEA}	Ψ_2^{DEA}	Ψ^{DEA}	Ψ_1^{CEAT}	Ψ_2^{CEAT}	Ψ^{CEAT}
One pollution-generating input	2014	0.7826	0.6860	0.5003	2.2207	0.7256	0.6672
	2015	12.9004	0.6621	0.7271	14.8408	0.7091	0.7979
	2016	13.4047	0.5758	0.6794	15.8236	0.6251	0.7573
	2017	1.2288	0.5404	0.4793	2.9480	0.5996	0.6228
	2018	0.6072	0.6128	0.4473	2.2083	0.6659	0.6331
	2019	0.9104	0.6167	0.4847	2.2149	0.6600	0.6296
	2020	0.9743	0.6481	0.4984	2.9146	0.7050	0.6638
	2021	0.9631	0.6487	0.4981	3.4833	0.6970	0.6825
	2014-2021	3.9715	0.6238	0.5393	5.8318	0.6734	0.6818
Two pollution-generating inputs	2014	0.7826	0.6384	0.4765	2.2207	0.7159	0.6624
	2015	12.9004	0.6195	0.7058	14.8408	0.6913	0.7890
	2016	13.4047	0.5557	0.6694	15.8236	0.6150	0.7523
	2017	1.2288	0.5205	0.4693	2.9480	0.5918	0.6189
	2018	0.6072	0.5844	0.4330	2.2083	0.6508	0.6255
	2019	0.9104	0.5987	0.4757	2.2149	0.6577	0.6285
	2020	0.9743	0.6159	0.4823	2.9146	0.6898	0.6561
	2021	0.9631	0.6180	0.4828	3.4833	0.6855	0.6767
	2014-2021	3.9715	0.5939	0.5243	5.8318	0.6622	0.6762

The results reveal notable differences in the efficiency patterns captured by the two estimators across both sub-technologies and the two pollution-generating input specifications. In the case of DEA, the mean inefficiency scores for $\boldsymbol{\Psi}_1$ exhibit considerable year-to-year fluctuations, with pronounced peaks in 2015 and 2016 (reaching 12.9004 and 13.4047, respectively, under the single pollution-generating input specification). These extreme values reflect the limited discriminating power of conventional DEA in certain years, likely attributable to the convexity assumption and the resulting overfitting of the production frontier. Excluding these outlier years, the mean scores for $\boldsymbol{\Psi}_1$ range from 0.6072 to 1.2288, with no clear monotonic trend over the study period. Conversely, the mean inefficiency scores for $\boldsymbol{\Psi}_2$ demonstrate a gradual decline from 2014 to 2017 (decreasing from 0.6860 to 0.5404), followed by a modest increase in subsequent years, stabilizing around 0.6487 by 2021. This pattern suggests that urban road systems achieved some progress in mitigating undesirable outputs during the mid-2010s, although this improvement is not sustained. Notably, the mean inefficiencies in $\boldsymbol{\Psi}_2$ consistently exceed those in $\boldsymbol{\Psi}_1$ (excluding the 2015-2016 anomalies), indicating that the potential for reducing undesirable outputs is greater than that for expanding desirable outputs under the DEA estimator.

Turning to the CEAT results, a markedly different pattern emerges. First, the mean inefficiency scores for $\boldsymbol{\Psi}_1$ are substantially higher than their DEA counterparts across all years, ranging from

2.2083 to 15.8236 under the single pollution-generating input specification. This elevation reflects the enhanced discriminating power of the CEAT method, which, by constructing a convexified frontier from regression tree partitions, identifies larger inefficiencies than the traditional DEA approach. Second, unlike the volatile DEA estimates, the CEAT scores for $\boldsymbol{\Psi}_1$ follow a more discernible temporal pattern: they increase sharply from 2014 to 2016, decline to a trough in 2018-2019, and subsequently rise again toward the end of the period. This U-shaped trajectory suggests that inefficiency in desirable output production first widened, then narrowed, before expanding once more. Third, the mean inefficiency scores for $\boldsymbol{\Psi}_2$ under CEAT exhibit greater stability, fluctuating within a narrow range of 0.5996 to 0.7256, with a slight upward trend in the later years. Consequently, the combined $\boldsymbol{\Psi}$ scores display greater year-to-year variation than those derived from DEA, reflecting its capacity to capture temporal dynamics in environmental inefficiency.

Comparing the two estimators, CEAT consistently yields higher mean inefficiency scores than DEA for both sub-technologies throughout the entire period, with the divergence being particularly pronounced for $\boldsymbol{\Psi}_1$. This finding aligns with the methodological properties of CEAT, which, by convexifying the EAT-based frontier, mitigates the overfitting problem inherent in DEA while preserving the flexibility of tree-based partitions. Furthermore, a noteworthy contrast emerges between the two methods regarding the relative magnitudes of $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ inefficiencies. In the DEA results, the mean scores for $\boldsymbol{\Psi}_2$ exceed those for $\boldsymbol{\Psi}_1$ in most years, indicating greater scope for abating undesirable outputs. However, in the CEAT results this pattern is reversed: the mean inefficiencies for $\boldsymbol{\Psi}_1$ are substantially larger than those for $\boldsymbol{\Psi}_2$ across all years, suggesting that when the frontier is estimated with greater precision, urban road traffic systems exhibit far more potential for expanding desirable outputs than for reducing undesirable ones.

Regarding the two pollution-generating inputs specification, the observed patterns are broadly consistent across both estimators. The inclusion of street lights as an additional pollution-generating input yields marginally lower mean inefficiency scores for $\boldsymbol{\Psi}_2$ in most years, reflecting the reclassification of this infrastructure element from a non-polluting to a pollution-generating role. This adjustment slightly alters the benchmark against which undesirable output performance is evaluated. Nonetheless, the overarching temporal trends and methodological contrasts remain robust to this specification change.

Synthesizing across estimators, a methodological continuum emerges. At one extreme, DEA and FDH, despite their differing convexity assumption, converge on the conclusion that inefficiency in bad output reduction ($\boldsymbol{\Psi}_2$) outweighs that in good output expansion ($\boldsymbol{\Psi}_1$). At the other extreme, CEAT amplifies the opposite hierarchy, with $\boldsymbol{\Psi}_1$ inefficiency dominating. Positioned intermediately, EAT captures the qualitative pattern of CEAT ($\boldsymbol{\Psi}_1 > \boldsymbol{\Psi}_2$) but with magnitudes more tempered, suggesting that the tree-based partitioning alone (without the subsequent convexification step) is sufficient to reverse the efficiency ranking observed under DEA and FDH. These findings underscore the pivotal role of estimator choice in operational efficiency analysis: while nonconvex estimators validate the convexity-related concerns raised by DEA, the machine learning-informed partitions of EAT and CEAT reveal that urban road systems possess considerably more untapped potential for expanding desirable outputs than for reducing undesirable ones, a distinction with profound implications for policy prioritization. Consequently, these results motivate continued exploration of semi- and nonparametric frontiers to fully characterize the robustness of these patterns.

5.3 Results of the Li-test and KSW-test for By-production DDF

This part lists the results for two tests regarding the convexity of the technology against a non-convex alternative. First, these efficiency measures are compared by means of a nonparametric test comparing two entire distributions as initially developed by Li (1996) and refined by Fan and Ullah (1999) and by Li, Maasoumi, and Racine (2009). The Li-test statistic tests for the eventual significance of differences between two kernel-based estimates of density functions f and g of a random variable x. The null hypothesis states that both density functions are almost everywhere equal ($H_0 : f(x) = g(x)$ for all x). The alternative hypothesis negates this equality of both density functions ($H_1 : f(x) \neq g(x)$ for some x). This test is valid for both dependent and independent variables.

Table 4 presents the results of the Li-test evaluating whether the distributions of efficiency scores derived from alternative by-production technologies differ significantly. This test compares the kernel-based density estimates of the efficiency distributions obtained from pairs of methods, specifically EAT versus CEAT and FDH versus DEA, for the sub-technologies $\boldsymbol{\Psi}_1$ (desirable output production), $\boldsymbol{\Psi}_2$ (undesirable output abatement), and for the aggregate technology $\boldsymbol{\Psi}$ as well. The analysis is conducted annually from 2014 to 2021, as well as over the whole period, under both single and two pollution-generating input specifications. Every p-value obtained is lower than 0.05, which allows us to conclude that the proposed models are significantly different from one another at the critical 5% level.

Under the single pollution-generating input specification, the test results reveal that the efficiency distributions for $\boldsymbol{\Psi}_1$ derived from EAT and CEAT are statistically different across all years, with p-values below 0.01. Similarly, the comparison between FDH and DEA for $\boldsymbol{\Psi}_1$ yields consistently significant differences, underscoring the sensitivity of efficiency measurement in desirable output production to the underlying methodological assumptions, particularly regarding convexity and frontier construction.

For $\boldsymbol{\Psi}_2$, the differences between EAT and CEAT are less pronounced in several years (e.g., 2017, 2018, 2020, and 2021), where the p-values exceed conventional significance thresholds. Likewise, the FDH versus DEA comparison for $\boldsymbol{\Psi}_2$ fails to reject the null hypothesis in most years. These findings suggest that the estimation of inefficiency in undesirable output reduction is relatively more robust to methodological variation, potentially due to the structural constraints imposed by pollution-generating inputs and the nature of bad output regulation.

However, when considering the aggregate technology $\boldsymbol{\Psi}$ all pairwise comparisons (EAT versus CEAT, and FDH versus DEA) exhibit highly significant differences across every year and for the full sample period ($p < 0.01$). This indicates that, despite occasional similarities in sub-technology estimates, the overall inefficiency distributions are strongly shaped by the choice of estimator,

Table 4: Result of the Li-test for by-production DDF comparing EAT vs. CEAT and FDH vs. DEA per year

Case	Period	Test	Ψ_1^{EAT} vs. Ψ_1^{CEAT}	Ψ_1^{FDH} vs. Ψ_1^{DEA}	Ψ_2^{EAT} vs. Ψ_2^{CEAT}	Ψ_2^{FDH} vs. Ψ_2^{DEA}	Ψ^{EAT} vs. Ψ^{CEAT}	Ψ^{FDH} vs. Ψ^{DEA}
One pollution-generating input	2014	Li-test	18.7949***	12.2282***	2.4699***	0.8864	26.3291***	13.9101***
		p-values	0.0000	0.0000	0.0160	0.1680	0.0000	0.0000
	2015	Li-test	34.9890***	34.2049***	2.8710***	0.3863	27.2714***	23.2002***
		p-values	0.0000	0.0000	0.0030	0.3090	0.0000	0.0000
	2016	Li-test	33.3622***	33.6505***	2.1654***	1.8340***	22.5856***	14.2509***
		p-values	0.0000	0.0000	0.0240	0.0490	0.0000	0.0000
	2017	Li-test	11.7957***	23.7791***	0.7976	0.9932	19.7780***	6.4359***
		p-values	0.0000	0.0000	0.1970	0.1390	0.0000	0.0000
	2018	Li-test	6.4359***	16.7104***	0.7649	0.1335	26.2848***	3.2875***
		p-values	0.0000	0.0000	0.1930	0.3910	0.0000	0.0020
	2019	Li-test	19.2125***	12.8737***	2.8691***	1.2609**	15.3114***	4.1713***
		p-values	0.0000	0.0000	0.0100	0.0850	0.0000	0.0020
	2020	Li-test	12.2346***	13.4579***	-0.5577	1.3055**	16.2884***	8.9896***
		p-values	0.0000	0.0000	0.6920	0.0860	0.0000	0.0000
	2021	Li-test	15.0176***	16.2636***	0.8276	1.4194**	16.5381***	6.3436***
		p-values	0.0000	0.0000	0.1680	0.0810	0.0000	0.0000
	2014-2021	Li-test	49.2942***	146.9439***	15.8703***	12.7547***	147.7700***	40.1391***
		p-values	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Two pollution-generating inputs	2014	Li-test	18.7949***	12.2282***	3.7644***	0.5626	26.3106***	13.9389***
		p-values	0.0000	0.0000	0.0000	0.2400	0.0000	0.0000
	2015	Li-test	34.9890***	34.2049***	4.1946***	1.9780***	29.2205***	10.3122***
		p-values	0.0000	0.0000	0.0000	0.0360	0.0000	0.0000
	2016	Li-test	33.3622***	33.6506***	2.1884***	3.4926***	22.5793***	14.8793***
		p-values	0.0000	0.0000	0.0250	0.0050	0.0000	0.0000
	2017	Li-test	11.7957***	23.7791***	1.9370***	0.7908	20.6767***	7.1002***
		p-values	0.0000	0.0000	0.038	0.156	0.0000	0.0000
	2018	Li-test	20.6724***	16.7104***	2.2746***	3.7011***	26.4168***	7.1230***
		p-values	0.0000	0.0000	0.0240	0.0010	0.0000	0.0000
	2019	Li-test	19.2126***	12.8737***	4.0885***	2.0414***	21.3435***	5.4803***
		p-values	0.0000	0.0000	0.0020	0.0350	0.0000	0.0000
	2020	Li-test	12.2346***	13.4579***	0.7835	4.2391***	18.5196***	6.7253***
		p-values	0.0000	0.0000	0.1680	0.0000	0.0000	0.0000
	2021	Li-test	15.0176***	16.2636***	2.0663***	2.1647***	17.3507***	6.3330***
		p-values	0.0000	0.0000	0.0240	0.0270	0.0000	0.0000
	2014-2021	Li-test	131.4496***	146.9439***	25.8534***	23.3411***	128.8922***	49.2942***
		p-values	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

reinforcing the importance of methodological pluralism in environmental efficiency analysis.

Under the two pollution-generating inputs specification, this overall pattern remains consistent. The test statistics for $\mathbf{\Psi}$ continue to show strong significance across all comparisons. Notably, the significance levels for $\mathbf{\Psi}_2$ comparisons become more pronounced relative to the single-input case, suggesting that expanding the set of pollution-generating inputs sharpens the distinction between estimation methods in capturing abatement inefficiencies.

In summary, all p-values are below 0.05 and in most cases below 0.01 leading to a clear rejection of the null hypothesis that the efficiency distributions are equal across methods. These results provide robust statistical evidence that the choice of frontier estimation technique materially affects the measurement of operational inefficiency within the by-production framework. They further validate the methodological contribution of this study, demonstrating that ML approaches such as EAT and CEAT capture distinct dimensions of production inefficiency that are not fully reflected in conventional DEA or FDH estimators.

While the Li-test confirms that efficiency distributions differ across estimators, it does not test convexity directly. Therefore, we apply the convexity tests developed by Kneip, Simar, and Wilson (2016) and Simar and Wilson (2020) separately to $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$, since imposing convexity without testing is particularly problematic in the by-production framework given the distinct mechanisms underlying desirable and undesirable outputs. Thus, we also report the results of this specific convexity test to explicitly test the convexity of the (sub)technology against a nonconvex alternative. These convexity tests compare the mean of the nonconvex and convex input-oriented efficiency measures denoted μ_i^{NC} and μ_i^{C}, respectively. The null hypothesis indicates that these two means are equal ($H_0 : \mu_i^{NC} = \mu_i^{C}$). The alternative hypothesis negates this equality and considers the nonconvex mean of to be larger than the convex mean ($H_1 : \mu_i^{NC} > \mu_i^{C}$). Since these tests assume independent sample means, this involves randomly splitting the original sample into two independent parts. The resulting one-sided test statistic is asymptotically normally distributed. Since different results are obtained from different sample splits, Simar and Wilson (2020) suggest multiple splits and employing bootstrap methods to implement tests based on sample means of the test statistics from each split, or based on the p-values obtained on each split. This yields two tests: one based on the mean of test statistics over multiple sample-splits (denoted KSW#1), and the other based on a Kolmogorov-Smirnov test of the uniformity of the resulting p-values (denoted KSW#2). In particular, we use 100 sample-splits as suggested by Simar and Wilson (2020), 1000 bootstraps to compute the sampling distribution of these two tests under the null hypothesis of convexity, and 100 bootstraps to compute the bias term. For more details about these two tests: see Simar and Wilson (2020).

More recently, Simar and Wilson (2026) propose a computationally efficient procedure for combining test results obtained from multiple sample splits through a simple decision rule. They show that the resulting test controls the asymptotic size, thereby ensuring that it does not exceed the prescribed nominal level. The test outcomes proposed by Simar and Wilson (2026) can be summarized using three dummy variables: SW10, SW5, and SW1. Each of these dummies is equal to

1 if the null hypothesis of the convexity is rejected at the 0.10, 0.05, or 0.01 significance level, respectively, and 0 otherwise.[1]

Given the distinct functional roles of the two sub-technologies within the by-production framework, we conduct the convexity test separately for $\boldsymbol{\Psi}_1$ (technology governing desirable output production) and $\boldsymbol{\Psi}_2$ (technology governing undesirable output abatement). In line with the directional orientation of the inefficiency measures employed in Sections 5.1 and 5.2, we evaluate $\boldsymbol{\Psi}_1$ using an output-oriented efficiency metric and $\boldsymbol{\Psi}_2$ using an input-oriented metric. The results are reported in Tables 5 and 6, in the latter case for both the single and two pollution-generating input specifications.

Table 5: Convexity Tests for Technology 1 ($\boldsymbol{\Psi}_1$)

Period	KSW#1	p-values	KSW#2	p-values	SW10	SW5	SW1
2014	-0.077	0.666	0.142	0.768	1.000	1.000	0.000
2015	-2.883	1.000	0.773	0.000	0.000	1.000	1.000
2016	-2.603	1.000	0.795	0.000	0.000	1.000	1.000
2017	0.282	0.115	0.188	0.371	1.000	1.000	1.000
2018	0.632	0.080	0.290	0.061	1.000	1.000	1.000
2019	-0.503	0.893	0.348	0.004	0.000	1.000	1.000
2020	-0.899	0.994	0.370	0.002	0.000	0.000	1.000
2021	-1.483	1.000	0.499	0.000	0.000	0.000	0.000

The results in Table 5 provide compelling evidence against the convexity of $\boldsymbol{\Psi}_1$. While the KSW#1 statistic rejects convexity only in 2018, the KSW#2 statistic rejects convexity in six out of the eight years. Moreover, the new Simar and Wilson (2026) test indicates that evidence against convexity is found in all years except 2021.[2] The overwhelming evidence supports the adoption of nonconvex estimators, such as FDH or EAT, for modeling desirable output production in urban road systems. The convexity violation implies that imposing a convex hull on $\boldsymbol{\Psi}_1$ may artificially smooth the frontier, thereby concealing genuine heterogeneity in production possibilities and distorting efficiency measurements.

Table 6 presents the convexity test results for the technology capturing the generation and abatement of undesirable outputs $\boldsymbol{\Psi}_2$. Similar to $\boldsymbol{\Psi}_1$, there is substantial evidence against the null hypothesis of convexity, although the evidence is slightly weaker for $\boldsymbol{\Psi}_2$ than for $\boldsymbol{\Psi}_1$. In the model with one pollution-generating input, the KSW#1 statistic does not reject convexity in any year, while the KSW#2 statistic rejects convexity in five out of the eight years. In addition, the new Simar and Wilson (2026) test provides some evidence against convexity in three of the eight years. In the model with two pollution-generating inputs, the KSW#1 statistic again fails to reject convexity in every year, whereas the KSW#2 statistic rejects convexity in all years. Moreover, the

[1] See the command *test.convexity.fast* in the *FEAR* package for *R* for more details.

[2] Note that under the new tests proposed by Simar and Wilson (2026), it is possible for SW10 =0 while SW5 =1 (as well as other seemingly inconsistent combinations); that is, the null hypothesis of convexity is rejected at the 5% significance level but not at the 10% significance level. See Simar and Wilson (2026) for further details.

new Simar and Wilson (2026) test indicates some evidence against convexity in six of the eight years. The evidence in Table 6 supports the adoption of nonconvex estimators, such as FDH or EAT, for modeling undesirable output production in urban road systems. The convexity violation implies that imposing a convex hull on $\boldsymbol{\Psi}_2$ may artificially smooth the frontier, thereby concealing genuine heterogeneity in production possibilities and distorting efficiency measurements.[3]

Table 6: Convexity Tests for Technology 2 ($\boldsymbol{\Psi}_2$)

Case	Period	KSW#1	p-values	KSW#2	p-values	SW10	SW5	SW1
One pollution-generating input	2014	-1.775	0.855	0.528	0.146	0.000	0.000	0.000
	2015	-1.727	0.904	0.517	0.103	0.000	0.000	1.000
	2016	-1.621	0.997	0.529	0.001	0.000	0.000	1.000
	2017	-1.365	0.992	0.430	0.009	0.000	0.000	1.000
	2018	-1.876	0.997	0.564	0.002	0.000	0.000	0.000
	2019	-1.720	1.000	0.561	0.000	0.000	0.000	0.000
	2020	-1.656	0.952	0.514	0.047	0.000	0.000	0.000
	2021	-1.361	0.939	0.421	0.105	0.000	0.000	0.000
Two pollution-generating inputs	2014	-1.472	0.873	0.505	0.079	0.000	0.000	1.000
	2015	-1.587	0.935	0.487	0.087	0.000	0.000	1.000
	2016	-1.747	0.999	0.553	0.001	0.000	0.000	1.000
	2017	-1.692	1.000	0.473	0.006	0.000	0.000	1.000
	2018	-1.742	0.994	0.528	0.010	0.000	0.000	1.000
	2019	-2.197	1.000	0.675	0.000	0.000	0.000	1.000
	2020	-1.771	0.995	0.572	0.004	0.000	0.000	0.000
	2021	-1.549	0.993	0.548	0.003	0.000	0.000	0.000

The slightly divergent convexity properties of $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ carry important methodological and policy implications. First, they validate the by-production framework's premise that distinct sub-technologies within a complex system may exhibit fundamentally different geometric structures. Second, they eventually provide empirical justification for a hybrid estimation strategy: while the convexity assumption may or may not be maintained for the abatement technology, it is rejected for the desirable-output technology, reinforcing the need for flexible nonconvex alternatives such as EAT and CEAT. Finally, from a policy perspective, the eventual convexity of $\boldsymbol{\Psi}_2$ suggests that marginal abatement efforts follow predictable, diminishing-returns patterns (which is not the case under nonconvexity), whereas the nonconvex nature of $\boldsymbol{\Psi}_1$ implies that improvements in service output efficiency may require more radical, discontinuous innovations rather than incremental adjustments.

In summary, the KSW convexity tests and the new Simar and Wilson (2026) test complement and reinforce the distributional comparisons obtained from the Li-test above. Together, they establish that (i) the two sub-technologies within the by-production framework possess somewhat distinct geometric properties; (ii) ML-based estimators like EAT and CEAT capture dimensions of inefficiency that are either obscured or distorted under conventional convexity assumptions; and (iii) methodological pluralism is essential for obtaining a comprehensive and policy-relevant portrait of urban road system performance.

[3]It is worth noting that both the KSW tests and the new Simar and Wilson (2026) test require large sample sizes and relatively low dimensionality, as measured by the number of inputs and outputs. Consequently, our test may also be affected by small sample sizes and high dimensionality.

6 Conclusions and Future Work

Following Guillen, Aparicio, Kapelko, and Esteve (2025), this paper departs from conventional studies that rely predominantly on either convex or nonconvex methodologies for operational efficiency assessment. Instead, it introduces Efficiency Analysis Trees (EAT) and its convexified extension (CEAT) to measure operational inefficiency in urban road traffic systems. By integrating these tree-based approaches with a by-production framework, we independently estimate the operational efficiency of two constituent sub-technologies: one for generating desirable outputs, and another for mitigating undesirable outputs. Using a panel dataset of 119 Chinese cities from 2014 to 2021, we systematically compare four frontier estimation methods in evaluating the operational efficiency of these systems: FDH, DEA, EAT, and CEAT.

Our empirical analysis yields several critical findings. First, efficiency estimates for both desirable and undesirable outputs vary significantly across methods, as confirmed by Li-test results showing persistent and statistically significant distributional differences between EAT and CEAT, as well as between FDH and DEA for most years. This underscores the substantial impact of methodological choice on efficiency outcomes.

Second, the KSW convexity tests reveal a clear asymmetry in the geometric properties of the two sub-technologies. For the desirable-output technology ($\mathbf{\Psi}_1$), convexity is consistently rejected across most years and test specifications, providing strong ex-post validation for using nonconvex estimators such as FDH or EAT. For the undesirable-output technology ($\mathbf{\Psi}_2$), the evidence is more mixed: the KSW#1 test (based on sub-sampling) fails to reject convexity in all years, while the KSW#2 test (based on bootstrapping) rejects convexity in five out of eight years under the one polluting-generation input specification, and the SW test also indicates violations in several years, with similar results prevailing for the two polluting-generation inputs specification. Thus, while convexity violations are more pronounced and systematic for $\mathbf{\Psi}_2$, the undesirable-output technology can be declared either convex or nonconvex depending on the test. These findings provide the first statistical foundation for a hybrid estimation strategy within the by-production framework, applying nonconvex methods to the desirable-output technology while possibly retaining convex methods for the undesirable-output technology.

Third, regarding the sources of inefficiency, FDH and DEA consistently indicate that inefficiency associated with undesirable outputs exceeds that of desirable outputs, highlighting significant potential for pollution control and accident reduction. By contrast, EAT and CEAT reveal nontrivial inefficiencies in desirable-output production, occasionally surpassing those of undesirable outputs. This suggests that tree-based methods uncover hidden inefficiencies masked by traditional approaches. Crucially, the application of rigorous convexity tests reveals slight differences between the two sub-technologies: while the convexity assumption for the desirable-output technology ($\mathbf{\Psi}_1$) is consistently rejected, it is somewhat less rejected for the undesirable-output technology ($\mathbf{\Psi}_2$). This finding provides strong ex-post statistical validation for employing nonconvex estimators like EAT and FDH for $\mathbf{\Psi}_1$. If we had resorted to convex estimators without this test, we would have mistakenly assumed a technology that does not align with the data, thereby masking the substantial potential for service expansion revealed by our tree-based methods. Sensitivity analyses further confirm that these efficiency patterns remain robust, regardless of whether road lighting is treated as a non-polluting or pollution-generating input. Collectively, this study not only introduces novel methodological tools for assessing urban road transport efficiency, but it also offers empirical insights to support more sustainable urban transportation governance.

Our work is a strong warning that when using the by-production technology, one always has to test for convexity both when using FDH and DEA frontiers, and when using EAT and CEAT ML models. Indeed, far too often by-production empirical applications are published without even mentioning the issue of convexity testing (e.g., Ray, Mukherjee, and Venkatesh (2018)). But, also recent ML articles sometimes use CEAT without testing for the appropriateness of its convexity assumption: see, e.g., Aparicio, Zofío, Noonan, Thompson, and Woronkowicz (2025) or Gaganis, Pasiouras, Tasiou, and Zopounidis (2026).

Future research can build upon this work in several key areas. First, the methodology should be extended to other critical infrastructure domains, including water and energy systems. Second, adopting a spatio-temporal modeling framework would allow for a deeper investigation into the determinants of efficiency dynamics. Finally, the integration of inherently interpretable ML models is crucial for improving the explainability of the results and ensuring their practical relevance for evidence-based policymaking.

Declaration of competing interest

The authors confirm the absence of any financial or personal ties with individuals or organizations that could have biased the findings of this study. They also affirm that they hold no commercial or associative interests that could create a conflict of interest regarding this work.

Acknowledgment

We are grateful to Paul Wilson for helpful comments. This paper is funded by the National Natural Science Foundation of China (72301145, 72401056).

References

Ang, F., K. Kerstens, and J. Sadeghi (2023): "Energy productivity and greenhouse gas emission intensity in Dutch dairy farms: A Hicks-Moorsteen by-production approach under non-convexity and convexity with equivalence results," *Journal of Agricultural Economics*, 74(2), 492–509.

Aparicio, J., M. Esteve, and M. Kapelko (2023): "Measuring dynamic inefficiency through machine learning techniques," *Expert Systems with Applications*, 228, 120417.

Aparicio, J., M. Esteve, J. J. Rodriguez-Sala, and J. L. Zofio (2021): "The estimation of productive efficiency through machine learning techniques: Efficiency analysis trees," in *Data-enabled Analytics: DEA for Big Data*, ed. by J. Zhu, and V. Charles, pp. 51–92. Springer, Cham.

Aparicio, J., M. Kapelko, and J. L. Zofio (2020): "The measurement of environmental economic inefficiency with pollution-generating technologies," *Resource and Energy Economics*, 62, 101185.

Aparicio, J., J. L. Zofío, D. Noonan, D. Thompson, and J. Woronkowicz (2025): "An innovative approach to efficiency measurement: Combining convexified efficiency analysis trees and data envelopment analysis to benchmark US museums," *Socio-Economic Planning Sciences*, 102, 102316.

Banker, R. D., A. Charnes, and W. W. Cooper (1984): "Some models for estimating technical and scale inefficiencies in data envelopment analysis," *Management Science*, 30(9), 1078–1092.

Bao, Q., Z. Zhai, and Y. Shen (2022): "Assessing road safety development in European countries: A cross-year comparative analysis of a safety performance index," *Applied Sciences*, 12(19), 9813.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone (2017): *Classification and Regression Trees*. Chapman and Hall/CRC.

Chambers, R. G., Y. Chung, and R. Färe (1998): "Profit, directional distance functions, and Nerlovian efficiency," *Journal of Optimization Theory and Applications*, 98(2), 351–364.

Chen, X., A. Guo, J. Zhu, F. Wang, and Y. He (2022): "Accessing performance of transport sector considering risks of climate change and traffic accidents: Joint bounded-adjusted measure and Luenberger decomposition," *Natural Hazards*, 111, 1–24.

Chen, X., X. Yi, Y. Yang, Y. Liu, and X. Qu (2026): "Measuring urban road transportation efficiency: A nonparametric slack-based analysis with Malmquist and Luenberger productivity indices," *Transportation Research*, 205A, 104845.

Cherchye, L., B. De Rock, D. Saelens, M. Verschelde, and B. Roets (2024): "Productive efficiency analysis with unobserved inputs: An application to endogenous automation in railway traffic management," *European Journal of Operational Research*, 313(2), 678–690.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring Labor Efficiency in Post Offices," in *The Performance of Public Enterprises: Concepts and Measurements*, ed. by M. Marchand, P. Pestieau, and H. Tulkens, pp. 243–268. North Holland, Amsterdam.

Dong, C., Y. Zhang, Z. Wang, J. Liu, and J. Zhang (2024): "The hybrid systems method integrating STAMP and HFACS for the causal analysis of the road traffic accident," *Ergonomics*, 67(7), 971–994.

Esteve, M., J. Aparicio, A. Rabasa, and J. J. Rodriguez-Sala (2020): "Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees," *Expert Systems with Applications*, 162, 113783.

Fan, Y., and A. Ullah (1999): "On Goodness-of-fit Tests for Weakly Dependent Processes Using Kernel Method," *Journal of Nonparametric Statistics*, 11(1), 337–360.

Farrell, M. (1959): "The Convexity Assumption in the Theory of Competitive Markets," *Journal of Political Economy*, 67(4), 377–391.

Farzipoor Saen, R., B. Karimi, and A. Fathi (2025): "Unleashing efficiency potential: The power of non-convex double frontiers in sustainable transportation supply chains," *Socio-Economic Planning Sciences*, 98, 102143.

Gaganis, C., F. Pasiouras, M. Tasiou, and C. Zopounidis (2026): "Beyond profit: Rethinking corporate efficiency frameworks through a sustainability lens," *European Journal of Operational Research*, 331(3), 941–959.

Grinerud, K., W. K. Aarseth, and R. Robertsen (2021): "Leadership strategies, management decisions and safety culture in road transport organizations," *Research in Transportation Business & Management*, 41, 100670.

Guillen, M. D., J. Aparicio, M. Kapelko, and M. Esteve (2025): "Measuring environmental inefficiency through machine learning: An approach based on efficiency analysis trees and by-production technology," *European Journal of Operational Research*, 321(2), 529–542.

Guillen, M. D., V. Charles, and J. Aparicio (2025): "Enhanced efficiency assessment in manufacturing: Leveraging machine learning for improved performance analysis," *Omega*, 134, 103300.

Hu, Z., M. Smirnova, Y. Zhang, N. Smirnov, and Z. Zhu (2021): "Estimation of travel time through a composite ring road by a viscoelastic traffic flow model," *Mathematics and Computers in Simulation*, 181, 501–521.

Joo, Y. (2025): "Comparative Efficiency Analysis of OECD Health Systems: FDH vs. Machine Learning Approaches with Efficiency Analysis Trees (EAT and RFEAT)," *Cost Effectiveness and Resource Allocation*, 23(1), 4.

Kang, L. (2024): "Assessing road safety performance in Chinese provinces: A comprehensive analysis of the past decade," *Research in Transportation Business & Management*, 54, 101133.

Kneip, A., L. Simar, and P. W. Wilson (2016): "Testing hypotheses in nonparametric models of production," *Journal of Business and Economic Statistics*, 34(3), 435–456.

Li, Q. (1996): "Nonparametric testing of closeness between two unknown distribution functions," *Econometric Reviews*, 15(3), 261–274.

Li, Q., E. Maasoumi, and J. S. Racine (2009): "A nonparametric test for equality of distributions with mixed categorical and continuous data," *Journal of Econometrics*, 148(2), 186–200.

Love, P. E., L. A. Ika, J. Matthews, and W. Fang (2020): "Curbing poor-quality in large-scale transport infrastructure projects," *IEEE Transactions on Engineering Management*, 69(6), 3171–3183.

Maziotis, A., R. Sala-Garrido, M. Mocholi-Arce, and M. Molinos-Senante (2023): "A comprehensive assessment of energy efficiency of wastewater treatment plants: An efficiency analysis tree approach," *Science of The Total Environment*, 885, 163539.

Michelaraki, E., M. Sekadakis, C. Katrakazas, A. Ziakopoulos, and G. Yannis (2023): "One year of COVID-19: Impacts on safe driving behavior and policy recommendations," *Journal of Safety Research*, 84, 41–60.

Murty, S., R. R. Russell, and S. B. Levkoff (2012): "On modeling pollution-generating technologies," *Journal of Environmental Economics and Management*, 64(1), 117–135.

Ray, S. C., K. Mukherjee, and A. Venkatesh (2018): "Nonparametric measures of efficiency in the presence of undesirable outputs: A by-production approach," *Empirical Economics*, 54(1), 31–65.

Sala-Garrido, R., M. Mocholi-Arce, M. Molinos-Senante, and A. Maziotis (2024): "Applying the Efficiency Analysis Tree Method for Enhanced Eco-Efficiency in Municipal Solid Waste Management: A Case Study of Chilean Municipalities," *Clean Technologies*, 6(4), 1565–1578.

Schwartz, N., R. Buliung, A. Daniel, and L. Rothman (2022): "Disability and pedestrian road traffic injury: A scoping review," *Health & Place*, 77, 102896.

Shephard, R. (1970): *Theory of Cost and Production Functions*. Princeton University Press, Princeton.

Simar, L., and P. W. Wilson (2020): "Hypothesis testing in nonparametric models of production using multiple sample splits," *Journal of Productivity Analysis*, 53, 287–303.

Simar, L., and P. W. Wilson (2026): "A Fast Method for Implementing Hypothesis Tests with Multiple Sample Splits in Nonparametric Models of Production," *Computational Economics*, 67, 3777–3813.

Tsionas, M. G. (2022): "Efficiency estimation using probabilistic regression trees with an application to Chilean manufacturing industries," *International Journal of Production Economics*, 249, 108492.

Yang, C., J. Jiang, J. Zhou, M. Hitosug, and Z. Wang (2023): "Traffic safety and public health in China-Past knowledge, current status, and future directions," *Accident Analysis & Prevention*, 192, 107272.

Zhang, Y., M. Smirnova, A. Bogdanova, Z. Zhu, and N. Smirnov (2018): "Travel time estimation by urgent-gentle class traffic flow model," *Transportation Research*, 113B, 121–142.

Zofio, J. L., J. Aparicio, J. Barbero, and J. M. Zabala-Iturriagagoitia (2024): "Benchmarking performance through efficiency analysis trees: Improvement strategies for colombian higher education institutions," *Socio-Economic Planning Sciences*, 92, 101845.

Appendices: Supplementary Material

Appendix A Correlation analysis between the potential inputs and the undesirable outputs

Given that our current dataset does not include direct measurements of emissions, such as CO2 or PM2.5, caution is required in classifying variables within the production framework. Following the by-production approach proposed by Murty, Russell, and Levkoff (2012), in the absence of direct emission data, a variable should be identified as a "pollution-generating input" based on its non-negative correlation with undesirable outputs along the production frontier. This statistical relationship is essential to maintain the logical consistency of the model. Accordingly, to identify the most appropriate indicators, we conduct a series of correlation analyses, including Spearman, Pearson, and Kendall tests, between the candidate inputs and the undesirable outputs. Detailed results and interpretations are provided in Table A1.

Table A1: Correlation results between the potential inputs and the undesirable outputs

	Spearman	Pearson	Kendall
Land for Road Transportation Facilities (Square Kilometer)	0.5106	0.515	0.3604
Expenditure for Transportation/10,000 yuan	0.4501	0.3497	0.3114
Road Length (kilometers)	0.1844	0.2964	0.1232
Number of Street Lights	0.5297	0.4151	0.3741
Number of vehicles owned by civilians (vehicles)	0.5670	0.5259	0.3984

The key findings from the correlation analysis in Table A1 reveal a distinct hierarchy among the candidate variables. The number of vehicles owned by civilians exhibits the strongest and most consistent correlation with undesirable outputs across all three statistical tests (Spearman: 0.5670; Pearson: 0.5259), statistically validating it as the most robust candidate for a pollution-generating input. The number of street lights follows as the second-most correlated variable (Spearman: 0.5297; Pearson: 0.4151). In contrast, road length demonstrates the weakest correlation, confirming its unsuitability for classification as a bad input.

Based on these empirical results and the methodological requirements of the by-production approach, we propose the following specification. As the primary recommendation, the number of civilian-owned vehicles should be definitively employed as the pollution-generating input. Should the modeling framework necessitate the inclusion of a secondary bad input, the number of street lights represents the next statistically valid option, given its correlation profile.